Linear Discriminant Analysis

ROBERT H. RIFFENBURGH[1] and CHARLES W. CLUNIES-ROSS[2]

THE PROBLEM to be considered here is that of identifying, or of classifying, an observed individual as being a member of one of two "populations." This problem arises in some form in most sciences. A recent example is the problem, associated with certain international tensions, of classifying salmon caught in the North Pacific fishery as having arisen from the Asiatic or American salmon populations.

The populations are to be considered as giving rise to observable individuals each of which may be (partially) characterized by a set of k measurements. The measurements of individuals from either population are distributed as if they were independent observations on a multivariate distribution of probability. These distributions are assumed to be multivariate normal, with known parameters, for each population.

1. Statement of the Problem

When an individual is misclassified, there may or may not be loss functions associated with the misclassification. For the problems of this paper explicit results are not obtainable for general loss functions; we shall assume loss functions to be constants. Let us designate as α the loss associated with misclassification of an individual from population I and as β the loss associated with misclassification of an individual from population II; $\alpha, \beta > 0$. Also, there is the question of whether or not anything is known about the mixed population from which the individual to be classified is drawn; in particular, whether or not there are known a priori probabilities, under a random drawing, that an individual belongs to either of the parent populations. Let us designate the prior probabilities as p for population I and $q = 1 - p$ for population II.

It follows that there are four levels of the classificatory problem to be considered:

(1.1) (a) with loss functions and prior probabilities
(1.2) (b) with prior probabilities only
(1.3) (c) with loss functions only
(1.4) (d) with neither

Misclassifications are undesirable; however, there are no adequate common units in which the "undesirability" can be measured for all of the above levels. At each level there are two quantities for which some form of joint minimization is desired, viz.:

(1.5) (a) $\alpha p P_I$, $\beta q P_{II}$
(1.6) (b) $p P_I$, $q P_{II}$
(1.7) (c) αP_I, βP_{II}
(1.8) (d) P_I, P_{II}

where P_I is the probability that a random individual of population I is classified as having arisen from II, and P_{II} is the probability that a random individual of II is classified as having arisen from I.

These four pairs of quantities will be referred to indiscriminately as "error quantities."

Now either error quantity of a pair may be reduced to zero, but not both jointly. Thus, joint minimization of the error quantities is, to a certain extent, arbitrary. While various specifications of joint minimization can be formulated, the more reasonable are those which have already been proposed elsewhere in the literature, viz.:

(i) joint minimization may be specified as that which minimizes the sum of error quantities; let us denote this criterion as "minisum";
(ii) joint minimization may be specified as that which minimizes the larger of the error quantities; let us denote this criterion as "minimax."

[1] Present address: Department of Mathematics, University of Hawaii. This paper is a portion of a dissertation submitted in partial fulfillment of the Ph.D. degree at the Virginia Polytechnic Institute; research was in part sponsored by the National Cancer Institute of the U.S. Public Health Service. Manuscript received June 8, 1959.

[2] Virginia Polytechnic Institute, Blacksburg, Virginia. Research was sponsored by the National Science Foundation under grant NSF-G-1858.

The first of these was introduced on level (a) by Brown (1950) and the second introduced on level (b) by Welch (1939). There has been more recent work on discriminant analysis, some of which is at levels similar to this treatment, but little seems applicable as the risk functions are not well defined.

Each of these specifications leads to the choice of one out of a family of quadratic discriminators. However, there are two related major difficulties: one is the determination of which member of the family is appropriate (for the minimax solution), and the other is that the integrals giving P_I and P_{II} cannot be evaluated explicitly (for either solution), and no tables are available for the resulting P_I and P_{II}.

If the variance-covariance matrices of the two populations are equal, the quadratic discriminator reduces to a linear discriminator; the integrals for P_I and P_{II} may then be reduced to the incomplete integral of the standard normal density. This is always true for any linear discriminator.

If we let A be a row vector of direction numbers, X be a row vector of variables (representing the possible measurements on the individual), c be a constant, and let primes denote transposition, then a linear discriminator may be written:

$$AX' = c. \tag{1.9}$$

We lose no generality if we number the populations such that the individual represented by X is classified into population I if $AX' < c$ and into population II if $AX' \geqslant c$.

Let (m_1, σ_1^2), (m_2, σ_2^2) be the mean and variance of AX' when X is distributed as in populations I, II, respectively. Then it follows, using an obvious notation, that:

$$P_I = \int_{-\infty}^{\frac{m_1 - c}{\sigma_1}} N(0,1)\, dx \tag{1.10}$$

$$P_{II} = \int_{-\infty}^{\frac{c - m_2}{\sigma_2}} N(0,1)\, dx \tag{1.11}$$

2. *The Appropriate Linear Function*

For the case when the distributions have identical variance-covariance matrices, the vector A is well known (see, for example, Fisher, 1936), being the inverse of this common matrix multiplied by the vector of difference means. When the variance-covariance matrices are not equal but are proportionate, then the corresponding A (using either of the matrices) is still optimum under both the minisum and minimax criteria.

In many fields the assumption of proportionate but not necessarily equal variance-covariance matrices is not unreasonable. This situation occurs, for example, in marine biology. The Hawaiian tunas *ahi* (*Neothunnus macropterus*) and *ahipahala* (*Thunnus alalunga*) are similar in most respects, but the *ahi* is a larger and more complex fish. If weight, fork length, lengths of second dorsal and anal fins, and the ratio of the length of the pectoral fin to the fork length (which varies inversely as the first four variables) are taken to be the variables, the population variance-covariance matrices for the *ahi* and *ahipahala* are (expected to be) proportional but unequal. Another example is cited in the literature, although only two variables were used. Mottley (1941) found that the variances and covariance for head and body measurements of trout (*Salmo gairdnerii kamloops*) stocked in two Canadian lakes were proportional.

The optimum A for general dispersion matrices is not easy to derive. This problem is considered in another paper by the authors (1960). The current paper considers optimum c for given A and thus in what follows it is only necessary to consider that A has been determined either by the methods mentioned above or arbitrarily.

3. *The Constant c for Minimized Error Quantities*

We lose no generality if we let $m_2 > m_1$ and $\sigma_2 \geqslant \sigma_1$. The designation of the population having the larger standard deviation as population II is arbitrary. We may then make a scale transformation of ± 1, whichever is necessary to obtain $m_2 > m_1$.

We now wish to obtain expressions for the constant c which will minimize the error quantities under the minisum and minimax criteria, respectively.

Consider $\alpha pP_I + \beta qP_{II}$.

$$(3.1) \qquad \frac{\partial(\alpha pP_I + \beta qP_{II})}{\partial c} = -\frac{\alpha p}{\sigma_1\sqrt{(2\pi)}} \exp\left\{-\frac{1}{2}\left(\frac{m_1-c}{\sigma_1}\right)^2\right\} + \frac{\beta q}{\sigma_2\sqrt{(2\pi)}} \exp\left\{-\frac{1}{2}\left(\frac{c-m_2}{\sigma_2}\right)^2\right\}$$

Equating the derivative to zero and rearranging, we obtain

$$(3.2) \qquad 2 \ln\frac{\beta q\sigma_1}{\alpha p\sigma_2} + \left(\frac{m_1-c}{\sigma_1}\right)^2 - \left(\frac{c-m_2}{\sigma_2}\right)^2 = 0$$

which is a quadratic in c with minisum c as roots:

$$(3.3) \qquad c(ms) = \frac{1}{\sigma_2^2-\sigma_1^2}\left\{\sigma_2^2 m_1 - \sigma_1^2 m_2 \pm \sigma_1\sigma_2\left[(m_2-m_1)^2 - 2(\sigma_2^2-\sigma_1^2)\ln\frac{\alpha p\sigma_2}{\beta q\sigma_1}\right]^{\frac{1}{2}}\right\}$$

Equation (3.3) has three possibilities:

(1) when there are no real roots,
(2) when no roots fall in (m_1,m_2), and
(3) when one and only one root falls in (m_1,m_2).

If a root should fall at one of m_1,m_2, this may be considered as a limiting case of situation (2). Situation (1) is trivial; all individuals are classified into one population. In situation (2), linear discrimination is not very helpful; quadratic discrimination is indicated. In these situations, possibly (depending on parameters) there is no discrimination which will be much of an improvement over the classification of all individuals into one population or a purely random classification. Thus, situation (3) will be considered in this paper.

When a root falls in (m_1,m_2), this is the root which minimizes $\alpha pP_I + \beta qP_{II}$, and is therefore the root desired. The other root maximizes $\alpha pP_I + \beta qP_{II}$ and therefore will not be used. Since σ_2 has been arranged to be greater than σ_1, and the smaller root is less than m_1, the positive square root is required. When $\sigma_1 = \sigma_2$, c(ms) is the root in (m_1,m_2); the other root is infinite.

Consider now the minimizing max (αpP_I, βqP_{II}). αpP_I and βqP_{II} are monotonic, decreasing and increasing respectively, in c; and, therefore, the desired c is located such that $\alpha pP_I = \beta qP_{II}$. An explicit result will not be found in general, since the integrals have not been evaluated explicitly. If $\alpha p = \beta q$, we have the integrals identical except for upper limits of integration, and $\alpha pP_I = \beta qP_{II}$ reduces to

$$(3.4) \qquad \frac{m_1-c}{\sigma_1} = \frac{c-m_2}{\sigma_2}$$

Solving, we obtain a minimax c:

$$(3.5) \qquad c(mx) = \frac{m_1\sigma_2 + m_2\sigma_1}{\sigma_2+\sigma_1}$$

It should be noted that if $\sigma_1 = \sigma_2$ and $\alpha p = \beta q$, both c(ms) and c(mx) reduce to a c dependent upon only the centroids,

$$(3.6) \qquad \frac{m_1+m_2}{2} = c(m)$$

This c(m) is the population analogue of the c introduced for samples by Barnard (1935) and Fisher (1936) and currently used in linear discriminant analysis.

4. A Discussion of Levels and c's

The results (3.3) and (3.5) apply for the case in which loss functions and prior probabilities are known, i.e., (1.1). When either or both of these quantities are unknown, corresponding to (1.2), (1.3), or (1.4), the corresponding error quantities considered are given by (1.6), (1.7), or (1.8) respectively. The results corresponding to (3.3) and (3.5) are obtained readily by the following substitutions in (3.3) and (3.5):

(1.2) "prior probabilities only": $\alpha = \beta = 1$
(1.3) "loss functions only": $p = q = 1$
(1.4) "neither": $\alpha = \beta = p = q = 1$.

For level (a), where both prior probabilities and loss functions are known, the risk may be measured and specified. If the total risk is to be minimized, then c(ms) is the appropriate constant. If the risk is to be minimized, subject to the restriction that risks from each source are to be equal, then c(mx) is the appropriate constant.

For level (b), where prior probabilities only are known, then c(ms) minimizes the conditional probability of misclassification. However, if classification is only part of the problem at

hand, then it may be desirable, in order to avoid bias in later stages, say, to minimize, subject to equalizing the probabilities of the two types of misclassification; here c(mx) is the appropriate constant.

For example, consider a merchandizing situation. If the problem is to allocate a limited shipment of goods to two branches of the same store, the same management suffers the loss from understocking either branch, and c(ms) is the appropriate constant to use in specifying the quantities of goods to go to each branch. On the other hand, if the problem is to equalize buyer-seller risk, as in the case of an independent mediator handling quality control, then c(mx) is the appropriate constant to use in specifying the acceptable level of quality.

For levels (c) and (d), the error quantities are in no sense absolute quantities. Here c(mx) will be the most reasonable constant to use, since under the minimax solution the expected numbers of misclassifications are equal for the two populations.

In practice, α, β, p, and q may or may not be well defined conceptually, but either way will often, perhaps usually, be unknown. Thus a comparison between discriminators using c(ms), c(mx), and c(m) at level (d) is appropriate.

5. *Comparison of Discriminators*

Introduction. The discriminators may be compared on the basis of our minisum and minimax criteria. Let us designate these criteria respectively in terms of the error quantities as

(i) $P_s = P_I + P_{II}$

(ii) $P_x = \max(P_I, P_{II})$.

In comparing discriminators, it can happen that either one has both criteria less than or equal to those of the other or this does not occur. If the former holds, then the discriminator with the smaller criteria may be said to be better than the other. This is true whether the discrimination is linear or not.

For the purposes of this paper, A has been taken to be a vector of constants. Thus, while linear discriminators are functions of both A and c, our comparison need be concerned only with varying c's. The restriction to level (d) together with the vector of constants, A, enables us to keep the number of parameters down to two for comparisons of the discriminators $AX' = c(ms)$, $AX' = c(mx)$, and $AX' = c(m)$. c(ms) and c(mx) are the c's derived for the two criteria; both reduce to c(m) in the special case of equal dispersion matrices. c(m) is the population analogue of the c used in practice and is easier to compute than are c(ms) and c(mx). Since c(mx) and c(ms) each minimize one criterion, the comparisons will be to find the conditions under which c(m) leads to both smaller P_x than does c(ms) and smaller P_s than does c(mx). When these conditions are satisfied then c(m) may be regarded as a compromise between c(ms) and c(mx).

The two essential parameters will be defined as

$$B = \sigma_2/\sigma_1 \tag{5.1}$$

$$C = \frac{m_2 - m_1}{\sigma_2 + \sigma_1} \tag{5.2}$$

It can be seen that $B \geqslant 1$ and $C > 0$. If results in B and C should be tabulated, the tables would be symmetric in log B, — log B, and in C, — C.

Condition for reasonable linear discrimination. Under certain conditions, linear discrimination does not yield good results; an example of this is the situation in which the centroids of the two populations are the same. Any description of the conditions necessary for linear discrimination to be able to lead to reasonable results must be, to some extent, arbitrary. Generally, the situations in which linear discrimination may be rejected are typified by no root of c(ms) being contained in (m_1, m_2).

At level (d) there are always two real solutions of (3.3). By restricting our interest to the range (m_1, m_2) it follows from considerations of monotonicity, continuity, and limiting behavior that a necessary and sufficient condition for the existence of a root of (3.3) in this range is

$$\frac{1}{\sqrt{(2\pi)}\sigma_1}\exp\left\{-\frac{1}{2}\left(\frac{m_2 - m_1}{\sigma_1}\right)^2\right\} < \frac{1}{\sqrt{(2\pi)}\sigma_2} \tag{5.3}$$

since the left and right sides of the inequality are the densities of populations I and II at m_2. (5.3) may be rewritten in terms of B and C as follows:

$$C^2 > 2(B+1)^{-2}\ln B \tag{5.4}$$

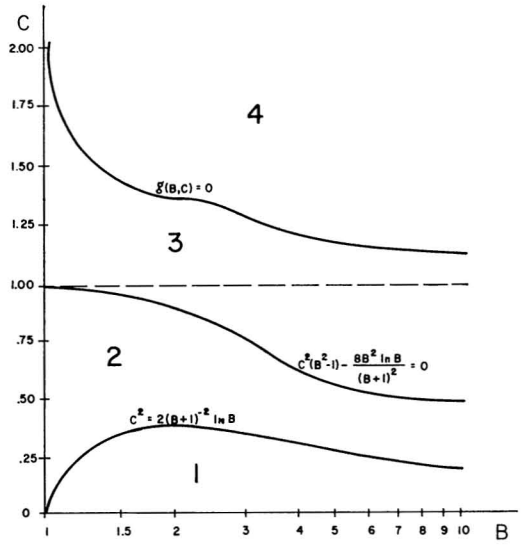

FIG. 1. Four regions in (B, C) corresponding to the properties: (1) no linear discriminator reasonable; (2) $c(m)$ is a compromise between $c(ms)$, $c(mx)$; (3) $c(ms)$ is better than $c(m)$; (4) both $c(mx)$, $c(ms)$ are better than $c(m)$. In general, the larger the C, the stronger will be the discriminator.

The lower curve in Figure 1 separates the regions in (B, C) for which (5.4) is true, untrue. Thus in region 1 a quadratic discriminator is appropriate; elsewhere a linear discriminator is appropriate.

Investigation of when c(ms) is better than c(m). Let us denote the larger conditional probability of misclassification, P_x, using c(m), c(ms) by $P_x(m)$, $P_x(ms)$ respectively.

Now c(mx) is the point on either side of which the probabilities of misclassification are equal, so that a $c < c(mx)$ indicates $P_I = P_x$ and a $c > c(mx)$ indicates $P_{II} = P_x$. Further, $m_2 \geqslant m_1$, $\sigma_2 \geqslant \sigma_1$ imply that both c(m) and c(ms) are greater than c(mx) since:

$$(5.5.a) \quad c(m)-c(mx) = \frac{(m_2 - m_1)(\sigma_2 - \sigma_1)}{2(\sigma_1 + \sigma_2)}$$

$$(5.5.b) \quad c(ms)-c(mx) = \frac{1}{\sigma_2^2-\sigma_1^2}\left\{m_1\sigma_2^2 - m_2\sigma_1^2 + \sigma_1\sigma_2\left[(m_2-m_1)^2 + 2(\sigma_2^2-\sigma_1^2)\ln\frac{\sigma_2}{\sigma_1}\right]^{\frac{1}{2}} - (\sigma_2-\sigma_1)(m_1\sigma_2 + m_2\sigma_1)\right\}$$

$$= \frac{\sigma_1\sigma_2}{\sigma_2^2-\sigma_1^2}\left\{m_2 - m_1 + \left[(m_2-m_1)^2 + 2(\sigma_2^2-\sigma_1^2)\ln\frac{\sigma_2}{\sigma_1}\right]^{\frac{1}{2}}\right\}$$

Therefore, $P_x(m) = P_{II}(m)$ and $P_x(ms) = P_{II}(ms)$.

It follows immediately that a necessary and sufficient condition for $P_x(m) > P_x(ms)$ is $c(m) > c(ms)$,

$$\text{i.e.,} \quad \frac{1}{2(\sigma_2^2-\sigma_1^2)}\left\{(m_1+m_2)(\sigma_2^2-\sigma_1^2) - 2(m_1\sigma_2^2 - m_2\sigma_1^2) - 2\sigma_1\sigma_2\left[(m_2-m_1)^2 + 2(\sigma_2^2-\sigma_1^2)\ln\frac{\sigma_2}{\sigma_1}\right]^{\frac{1}{2}}\right\} > 0$$

$$\text{i.e.,} \quad (m_2-m_1)(\sigma_2^2-\sigma_1^2) > 2\sigma_1\sigma_2\left[(m_2-m_1)^2 + 2(\sigma_2^2-\sigma_1^2)\ln\frac{\sigma_2}{\sigma_1}\right]^{\frac{1}{2}}$$

which may be rewritten as:

$$(5.6) \quad C^2(B^2-1) - \frac{8B^2\ln B}{(B+1)^2} > 0$$

The center curve in Figure 1 separates the regions of (B, C) for which P_x using $AX' = c(ms)$ is greater, less than those using $AX' = c(m)$. Thus in regions 1 and 2, c(m) is better with respect to the minimax criterion; in regions 3 and 4, c(ms) is better with respect to the minimax criterion.

Investigation of when c(mx) is better than c(m). Let us denote the sum of conditional probabilities of misclassification, P_s, using c(m), c(mx) by $P_s(m)$, $P_s(mx)$.

On expressing P_I, P_{II} in terms of c(m), c(mx) and hence in terms of B, C, it follows, after rearrangement, that

$$(5.7) \quad P_s(m)-P_s(mx) = 2\int_0^C N(0,1)dx - \int_0^{\frac{C(B+1)}{2}} N(0,1)dx - \int_0^{\frac{C(B+1)}{2B}} N(0,1)dx = g(B,C)$$

say. From differential-geometrical considerations and the fact that both c(m), c(ms) are greater than c(mx), it follows that $c(m) < c(ms)$ implies that $P_s(m) < P_s(mx)$. The upper curve in Figure 1 is the curve $g(B, C) = 0$, which separates the regions of (B, C) for which the sum of conditional probabilities of misclassification using $AX' = c(mx)$ is greater, less than those using $AX' = c(m)$. Thus in region 4, c(mx) is better with respect to the minisum criterion; elsewhere c(m) is better with respect to the minisum criterion. The asymptote as B tends to infinity is, approximately, $C = 1.029$.

Figure 2 shows g(B, C) plotted against C for several values of B.

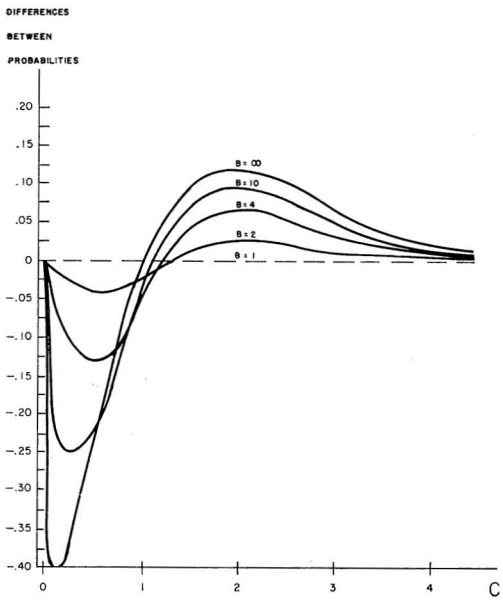

FIG. 2. Difference between (a) the sum of conditional probabilities of misclassification using $c(m)$, and (b) the same using $c(mx)$, expressed as a function of C for several values of B.

REFERENCES

BARNARD, M. M. 1935. The secular variations of skull characters in four series of Egyptian skulls. Am. Eugen. 6: 352–371.

BROWN, G. W. 1950. Basic principles for construction and application of discriminators. J. Clin. Psychol. 6: 58–61.

CLUNIES-ROSS, C. W., and R. H. RIFFENBURGH. 1960. Geometry and linear discrimination. Biometrika 47: in press.

FISHER, R. A. 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7: 179–188.

MOTTLEY, C. McC. 1941. The covariance method of comparing the head-lengths of trout from different environments. Copeia 3: 154–159.

SMITH, C. A. B. 1954. Biomathematics. Hafner, New York.

WELCH, B. L. 1939. Note on discriminant functions. Biometrika 31: 218–220.